DOCUMENT RESUME

ED 429 833 SE 062 396

AUTHOR Adkin, Sally; Halpin, Myra; Howe, Ann
TITLE A Non-Content Specific Test of High School Students' Progress in Science.
PUB DATE 1999-03-00
NOTE 12p.; Paper presented at the Annual Meeting of the National Association of Research in Science Teaching (Boston, MA, March 28-31, 1999).
PUB TYPE Reports - Research (143) -- Speeches/Meeting Papers (150) -- Tests/Questionnaires (160)
EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *High School Students; High Schools; *Problem Solving; *Science Education; Science Equipment; *Science Process Skills; Scientific Literacy; *Scientific Principles; *Skill Development
IDENTIFIERS *Science Achievement

ABSTRACT

This paper describes the development, administration, and results of an instrument to assess changes in students' science abilities as they progress from ninth grade through twelfth grade. Standard science tests commonly in use in schools to measure high school science students' achievement are content specific. Although these tests are useful they do not tell teachers or other educators what skills or general science knowledge students have acquired, nor can they ascertain students' progress as they move through high school. As part of the evaluation of the WINNERS II Project described in this paper, the research team wished to know more about what students were taking away from their science classes as they made their way through four years of high school. The research team decided to focus on designing a test that measured (1) understanding of the nature of science; (2) use of skills to solve problems; and (3) development of skills to use science equipment. (Contains 13 references and 3 tables.) (Author/WRM)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED 429 833

A Non-Content Specific Test of High School Students' Progress in Science

Sally Adkin & Myra Halpin*
North Carolina School of Science and Mathematics

Ann Howe, Educational Consultant, Adjunct
North Carolina State University

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☒ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

M. Halpin

*corresponding author
1219 Broad Street
Durham, NC 27705
(919) 286-3366
halpin@academic.ncssm.edu

BEST COPY AVAILABLE

SE060 3910

A Non-Content Specific Test of High School Students' Progress in Science

Purpose

This paper describes the development, administration and results of an instrument to assess changes in students' science abilities as they progress from ninth grade through twelfth grade. Standard science tests commonly in use in schools to measure high school science student's achievement are content specific. Although these test are useful they do not tell teachers or other educators what skills or general science knowledge students have acquired nor can you ascertain students' progress as they move through high school. As part of the evaluation of the WINNERS II Project described below, the team wished to know more about what students were taking away from their science classes as they made their way through the four years of high school. The team decided they would concentrate on designing a test that measured: 1) understanding the nature of science 2) using skills to solve problems; and 3) develop skills to use science equipment

History and Problem

WINNERS II was a three-year project administered by the North Carolina School of Science and Mathematics, NCSSM, and funded by the Glaxo Wellcome Foundation to work with high school science teachers to integrate technology into their curriculum. The project design was based on three cornerstones: professional development, technology infusion and updated science content. Professional development included summer workshops, on-site support by the North Carolina School of Science and Mathematics staff, experimentation with new curricula, and attendance and presentations at professional meetings. Project funds were used to purchase new lab materials, computers, multimedia hardware and software, and to train teachers to use these new tools. The school served, EWHS, was a 1500 student, rural, quickly becoming suburban high school, in North Carolina. The project staff worked with the 12-member science faculty, 2 special-programs teachers who teach science and the 2 media specialists.

As part of the evaluation design, the team wished to measure changes in students' science abilities as they progressed from ninth through twelfth grade. As the authors began their search for an appropriate testing instrument to measure secondary science understanding, they looked for commercially published standardized tests. Several of these exist at the grade 3-8 levels. Few if any seemed to exist at the high school level. Boston College's Center for the Study of Testing, Evaluation, and Educational Policy conducted a study of science and mathematics testing (Harmon, 1995). The study reviewed six standardized tests that dominate the school testing market in all fifty states: California Achievement Test, Comprehensive Test of Basic Skills, Iowa Test of Basic Skills, Survey of Basic Skills of Science Research Associates, Stanford Achievement Test, and Metropolitan Achievement Test. The Boston College study did not provide comparisons of general high school science understanding. At the high school level, each content was treated separately: Earth Science, Physical Science, Biology, and Chemistry. Physics was not included because enrollment makes up less that 5% of the high school population.

Next, test designers turned to leading professional organizations including the National Science Teachers Association (NSTA), American Association for the Advancement of Science (AAAS), and the National Research Council (NRC). These organizations have been leaders in science education reform. The AAAS has published a set of recommendations on what

understandings and habits of mind are essential for all citizens in a scientifically literate society (1989). NRC's new *National Science Education Standards* (1996) proposes changes in what students are taught, in how their performance is assessed, in teacher education, and the school's relationship with the community. NSTA in partnership with the National Association of Biology Teachers publishes a high school biology test. This test and test samples from the North Carolina Biology and Chemistry End of Course Tests were valuable guides for test developers. However, none of these organizations had one tool for measuring high school science understanding across content areas and grade levels.

The authors did find a number of middle school instruments, including *The Performance Process Skills Test (POPS), (*Pottenger, Mattheis, Jones, Nakaymama, 1988). The POPS, consisting of 21 multiple choice items, came close to what the NCSSM/ EWHS team hoped to accomplish with its secondary instrument. The emphasis was on scientific processes and emphasized higher order cognitive skills. Previous testing using this instrument with NCSSM students found that it did not discriminate with higher ability students.

After an exhaustive search no existing instrument was found that was non-content specific and appropriate for high school aged students. This report describes the design, piloting, and results of the instrument created to measure student's progress in science. The NCSSM/EWHS team sought to develop an instrument to assess students in grades 9 -12 science understanding. The resulting test differs from other secondary science standardized assessment tools, because it goes beyond any one content discipline and seeks to test the kind of science thinking most valued by recent science education reform efforts -content, skills, and application.

Development and Pilot Test of Instrument:

Science education reform literature suggest science instruction should emphasize a new way of teaching and learning about science that reflects the way science is actually done, emphasizing inquiry as a way of achieving knowledge and understanding about the world (NRC, 1996). The State of North Carolina has adopted five program goals that are the basis for scientific literacy for North Carolina's students. These are: 1) Understand the nature of science, 2) Become proficient in using science process skills to solve problems and make decisions, 3) Develop skills to manipulate and/or operate science equipment, 4) Develop responsible attitudes toward the environment, science technology, and science, 5) Understand the relevance of current topics in science (DPI, 1995). The NCSSM/EWHS team decided they would concentrate on designing a test that measured goals 1-3.

The following criteria were used to design the instrument.
The test should:

- be non-content specific so that student's score could be compared each year
- be authentic
- have a lab component
- measure the science skills of data gathering, analysis, and reporting
- be difficult enough to measure the brightest students
- written in such a way that even the special programs students could have some measure of success
- be one that teachers felt measured the skills they wanted each of their student to master
- be one that teachers felt ownership, and
- be administered in a 55 minute period.

The test was divided into three parts: (1) an open-ended graphing activity; (2) a laboratory practical; and (3) a multiple-choice test which included four questions from each of the major science disciplines.

In the graphing activity, students were given a set of data to graph and then asked to answer a series of questions using their graph. Students were to identify the dependent and independent variables and use logical intervals for the units on the axis. The questions required that they be able to make inferences from extrapolations of their graph. No instructions were given on extrapolation and students were instructed to explain the logic they used in reaching their solutions.

For the lab practical, students were given fifteen minutes at a lab station that contained a colored liquid, assorted glassware, and a balance. The students were asked to determine the density of the liquid and to answer a series of questions relating to the liquid. More equipment was provided than was necessary to solve the problem. The intent was to see if student could select the proper equipment and use it appropriately. Students were instructed to write down their procedure, the equipment they used and how it was used. After completing the lab activity, students were given the dimensions of a solid and asked if the solid would sink or float in the liquid and to explain their answer.

The multiple-choice portion of the test was composed of four questions from each science discipline. The questions were based on concepts the content experts on the NCSSM team believed all students should know when they graduate from high school. Although the test developers did not want the test to be content specific, questions were designed to reflect basic knowledge from the major disciplines. Teachers reviewed the selected questions and reworded them, as needed using "How to Write Multiple-choice Achievement Test Items" (NCDPI) as a guide.

The team also decided to add a series of questions on general experimental process. As these questions were developed, care was given to consider the cognitive level of each question. More questions would be designed at higher-order thinking levels. The cognitive taxonomy of Benjamin Bloom (1956) was used for categorizing cognitive levels. Figure 1 shows the item specification table for the test.

TABLE 1
Test Specifications for NCSSM/EWHS Multiple Choice Test
By Item Number

	Level			
Content	Knowledge/ Comprehension	Application/ Analysis	Evaluation/ Synthesis	Totals
Experimental Methods		1,2,3,4		4
Chemistry		7, 8	5,6,9	5
Physics		10	11,12,13	4
Biology	16,17	14,15	19	5
Earth Science	22	18, 21	20	4
Totals	3	11	8	22

Administration and Scoring

To test this three part instrument test developers choose a pool of 215 EWHS students in nine classes that reflect a population similar to the actual target population. Students in these classes represented a range typical of the population that would be tested the following spring and each spring of the project thereafter. In order to determine if the test would be useful for the brighter students, sixty students in an introductory physics class at NCSSM took the test. The test was administered in one 55-minute period under the same conditions as proposed for the official spring testing. For the multiple-choice section, students marked their answers on Scantron sheets that were mechanically scored.

The teachers all took the test then self corrected their answers using the multiple choice key and the graphing and lab rubrics provided by the Winner II team. The teachers and the Winners II team discussed the rubric and consensus was reached on how all papers would be scored. The scoring was rigorous by design to provide room for improvement for even the brightest students. The 20 member team scored the graph and lab portions following the revised rubric. The grading was conducted in teams of four. Two members graded the lab portion and two graded the graphing section. A set of papers were graded then exchanged with their partners for re-grading. The scores for each were compared and any differences were discussed. This procedure was repeated with the other half of the team to ensure all were using the rubric in the same manner. The four-member team worked together to resolve any difference in grading. The project director selected random papers from each group for comparison.

The following table lists the average score for each part of the test. Nine classes were tested in the pilot group and three in the NCSSM group.

Table 2 Results of the Pilot Testing of the Instruments

	Average Lab score	Average graph score	Average multiple test score	Total average
Pilot group n= 215	39%	59%	45%	44.9%
NCSSM n = 60	77%	85%	82%	80 %

These scores indicate that the instrument is challenging even for the brightest students. The following two sections will discuss an analysis of the data.

Psychometric Characteristics

Reliability of the Multiple Choice Test

Cronbach's Coefficient Alpha was measured at 0.6883, using "Statistical Production and Service Solutions" (SPSS 7.5). NCSSM's research office first examined item correlations and found Questions 9, 10 and 13 problematic for many students. After examining question 10 it was found that two possible choice (making the ramp longer would increase.. and making ramp shorter would decrease) even though both were wrong might have lead to some confusion. Question 13 correlations were almost all negative and item total correlation was negative and had a difficulty of .0654; about 6.54% of test takers getting this item correct. Cronbach's Alpha with items 9 and 13 removed was .7136 (This compares favorably with reliability data from the California Achievement Test, the Iowa Test of Basic Skills, and the Stanford Achievement Test. Reliability indices from these tests range .70 - .91 (Impara & Plake). The reliability index for the Middle School Science Test (*POPS*) was .75. At least 71.36% of the observed score variance

is attributable to true score variance for this examinee group. Fifty-one percent of the observed score variance on subsequent tests could be predicted by the variance observed on this first test. The correlation between observed scores and true scores is SQ Root (.7136) or about 0.84. (see below for additional discussion of problems 9 and 13)

Validity of Multiple Choice Test

A key issue for the content aspect of construct validity is assuring that the questions are relevant and representative of the domain. For this measure of the domain of secondary science ability, test authors drew on the expertise of the 20-Winners II project teachers. The team also compared content with other existing instruments. The EWHS teachers informally compared students' scores with grades.

–Principal Component analysis using SPSS 7.5 Total Variance Explained for all 22 items and with questions 9 and 13 removed found that with all 22 items, one factor explains 15.424% of the variance, with 9 and 13 removed, one factor explains 16.5% of the variance. Although these are not particularly impressive percentages, there is a big drop in percent of variance explained by a second factor. Three questions 9, 10, and 13 were answered incorrectly more than fifty per cent of the time. The problems encountered with questions 9,10, and 13 were no surprise to the WINNERS test constructors. Many of the 20 EWHS teachers who took the test in the review phase missed these questions. The team had discussed these items and was persuaded to include them, because it was believed to be a discriminator for the brightest students and also pointed out major misconceptions. A few minor changes were made with wording and diagrams clarified but none of the questions was removed.

In questions #9 mentioned above the students demonstrated a major misconception:

Heat is involved in a chemical reaction because

a. *chemical bonds are broken and others are formed*
b. *nuclear decay occurs*
c. *mass is converted into energy in the reaction*
d. *a phase change occurs*
e. *a bigger molecule is formed*

The correct answer is (a) but the most frequent student answer given was (c).

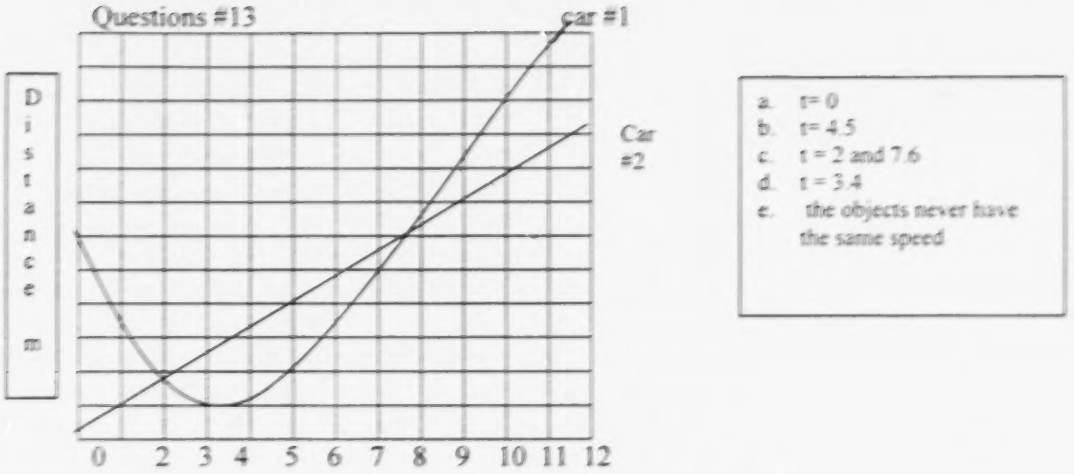
Question #10

A large block must be lifted from the floor to a shelf two meters above the floor. This can be accomplished by lifting the block straight up and setting it on the shelf or sliding it up a ramp (incline plane) to the shelf. In terms of work done on the block, and ***ignoring friction****,*

a. *more work is done on the block by lifting it straight up than by using the ramp*
b. *less work is done on the block by lifting it straight up than by using the ramp*
c. *more work would be needed if the ramp is made shorter*
d. *less work would be needed if the ramp is made longer*
e. *the work done on the block is the same for either method*

The correct answer is (e) but the most frequent answer given was (a). As you can see responses (c) and (d) could be confusing because they say the same thing.

In questions #13, students were asked to analyze data from a graph and determine the times when the two cars had the same speed. The correct answer (b) is the point at which the two lines are parallel. The most frequent answer was (c) the point at which the two lines intersect.



Dependability of Behavioral Measures - Lab and Graphing Activities

In order to examine the dependability of the two performance based measures, the team looked for potential sources of measurement error and sought to estimate the magnitude of such error according to generalizability theory (Cronbach, 1972). Test conditions, rater variance, and student performance were potential score facets. In the pilot study test conditions were essentially the same for all 215 students. Team leaders controlled for rater variance by rubric design, rater training, and cross-rater comparisons and random re-scoring.

Results of Administration of Instrument

The students who took the pilot test in the fall and also again in the spring showed little test retest change. In fact the average changes in the graph and lab activities had a slight negative change. The multiple choice showed a positive change. The following table shows the raw score and the percentage change for the students who participated in the pilot test and also repeated the test in the spring will all other students.

Table 3
Comparision of Student's Test Scores
N = 215

Average change Lab	Average change Graph	Average change Multiple choice	Total average change
-0.62	-0.27	1.03	.16
-2.3%	-1.9%	4.7%	.25%

After seeing the results, teachers of the students indicated no surprise in the results and reported that just prior to the pilot test most students had received instruction in collecting data and graphing data. These activities are included at the beginning of the year in many science courses. One teacher reports, "They probably forgot that from the beginning of the year."

The table below lists the average scores by grade level for all ability levels. In many of the lower ability classes several students answered few if any questions in the lab or graphing portions of the test. This brings the over-all averages down. We elected not to remove any scores because the primary objective is to see improvement for individual students, not just class averages.

Student Assessment for all East Wake High School Science Students
Table 4 --Spring 1997

Grade	Lab Average %	Graph Average %	Multiple Choice Average %	Total Average %
9	13	46	40	29
10	26	55	45	39
11	30	52	46	40
12	31	55	48	42
All Students	**23**	**51**	**44**	**36**

Table 5 --Spring 1998

Grade	Lab Average %	Graph Average %	Multiple Choice Average %	Total Average %
9	16	54	41	33
10	25	58	44	39
11	32	65	46	44
12	40	71	49	50
All Students	**26**	**60**	**44**	**40**

Several conclusions may be drawn from the assessment. First, there is a correlation between the test scores and grade level, as well as with class level. This indicates that the instrument can measure a change in a student's science skills and basic concept knowledge, however, this change may also be due to maturation. Secondly, many students did not know what to do to solve the lab practical, the authentic assessment in which students were given lab equipment and asked to solve a problem using the tools provided. Comments frequently heard were, "Where is the procedure sheet?", "I don't know what you want me to do," or "We didn't do this in my class." Some teachers scoring the assessment commented that they should probably give the students more opportunities to do this type of activity in order for them to know what to do on the test. Students were more successful in completing the graphing assessment but many failed to apply their knowledge from the graph to other questions. The students displayed a lack of self-confidence and an unwillingness to try the lab activity, most notably, the applied level students but this behavior was observed at all levels. Several of the advanced students expressed a great deal of frustration because they did not know "the right answer".

The scores listed above are averages of *all* students taking the tests each years. The table reflects the average change for *each* student who took the test both years.

Table 3 -- Matched Average Change Year 1 to Year 2

Change in Lab	Change in Graph	Change in Score	Change in Total
3.22	2.01	0.74	5.97
11.5%	14.4%	3.4%	9.3%

These data were obtained by taking the matched scores for each student who took the test both years. The differences in the scores for the student were averaged to show the change. There are dramatic gains in the lab and graphing portions of the test and small changes in the multiple choice science test.

Discussion of Results:

The project team feels the instrument has merit and gives insights to students' progress as well as, points out misconceptions. We also hypothesized that one fact which lead to improved scores may be due to teachers changing some lab activities from the traditional recipe type lab to a more open-end type experience. There also was an attempt by many teachers to provide more that one opportunity for students to practice skills and opportunities to go into the lab without a procedure sheet spelling out every step they should take in solving a problem.

To improve the instrument, an item analysis of individual multiple-choice questions should be correlated with the lab and graphing scores then refined as needed. The rubric and scoring procedures should be more carefully studied to ensure that the results do include all the possible correct responses that could have been logically used by students. For test administration beyond the pilot, an ANOVA estimation of variance components on the lab and graphing activities could provide a generalizability measure. Further analysis of individual item correlated with lab and graphing responses may also provide additional insights in to students' progress. This instrument should also be tested in other high school settings to determine it is indeed a good measure of student's progress in science.

Implications:

One of the important aspects of the development process is the involvement of teachers at every step in the development and administering of the instrument. This was not a test developed by "testing experts" but by experienced teachers. Involving the teachers in the development and grading of the test had several important impacts.

- Teachers valued the skills and concepts being tested.
- Teachers realized that many of their students did not know the basic concepts they assumed the students knew.
- Teachers recognized the need to provide more open-end types of experiences for students.
- Some teachers changed some of their labs from of the traditional varification model to more inquiry based exercises.
- Some teachers attempted to use more open-end type questions and fewer multiple-choice tests.

There is a great deal of discussion in the science education community about tests driving the curriculum. In North Carolina there are End –of-Course Exams for most courses on which teachers are evaluated. Teachers involved in the project reported that they did not mind being held accountable for what their students learn if the test actually reflected the goals they had for their students. The teachers involved in this project decided that having a test influence the curriculum is not necessarily a negative **if**, the test measures the desired outcome of good science education.

References

The Content Core. (1992). Washington, D. C.: National Science Teachers Association.

Bloom, B. (1956). *Taxonomy of educational objectives, handbook I: The cognitive domain*. New York: McKay.

Crocker L. and Algina J. (1986). *Introduction to Classical & Modern Test Theory*. Fort Worth: Harcourt Brace Jovanovich.

Cronbach, L.J., Gleser, G.C., Nanda, H. & Rajaratnam, N. (1972). The Dependability of Behavioral Measurements: Theory of Generalizability of Scores and Profiles. New York: Wiley.

Harmon M. (1995). The changing role of assessment in evaluating science education reform. In R. G. O'Sullivan (Ed.), *Emerging roles of evaluation in science education reform*. New Directions for Program Evaluation, No. 65, 31-51. San Francisco: Jossey-Bass.

How to Write Multiple-Choice Achievement Test Items. (1994). Raleigh: North Carolina Department of Public Instruction.

Impara, J.C., & Plake, B. S. (1998). *Burroughs 13th Mental Measurement Yearbook*. University of Nebraska Press.

Kulm, G. and Malcom S. (1991). *Science Assessment in the Service of Reform*. Washington, D.C.: American Association for the Advancement of Science.

Messick S. (1995). Validity of psychological assessment. *American Psychologist*. 50 (9). 741-749.

North Carolina Department of Public Instruction. (1995). *NC science program goals*. Raleigh, NC: DPI Publications.

Pottenger F., Mattheis F., Jones M., Nakayama G. (1988), *Development of the performance of process skills (POPS) test for middle grades students*. Greenville, NC: East Carolina State University.

Project 2061: Science for All Americans. (1989). Washington, DC: American Association for the Advancement of Science, Inc.

National Science Education Standards. (1996). National Research Council. Washington, DC: National Academy Press.